Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W243–W248***
doi:10.1093/nar/gkl298

BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences

Liangjiang Wang* and Susan J. Brown

Bioinformatics Center, Division of Biology, Kansas State University, Manhattan, Kansas 66506, USA

Received February 12, 2006; Revised March 1, 2006; Accepted April 7, 2006

ABSTRACT

BindN (http://bioinformatics.ksu.edu/bindn/) takes an amino acid sequence as input and predicts potential DNA or RNA-binding residues with support vector machines (SVMs). Protein datasets with known DNA or RNA-binding residues were selected from the Protein Data Bank (PDB), and SVM models were constructed using data instances encoded with three sequence features, including the side chain pK_a value, hydrophobicity index and molecular mass of an amino acid. The results suggest that DNA-binding residues can be predicted at 69.40% sensitivity and 70.47% specificity, while prediction of RNA-binding residues achieves 66.28% sensitivity and 69.84% specificity. When compared with previous studies, the SVM models appear to be more accurate and more efficient for online predictions. BindN provides a useful tool for understanding the function of DNA and RNA-binding proteins based on primary sequence data.

INTRODUCTION

Many proteins perform essential functions through interactions with DNA or RNA molecules. For example, transcription factors bind to specific DNA motifs in the promoters, resulting in activation or repression of transcription (1). Protein–RNA interactions play pivotal roles in both assembly and function of ribosomes (2) as well as eukaryotic spliceosomes (3). Thus, identification of the amino acid residues that recognize DNA or RNA is important for understanding a variety of biological processes.

Analysis of structural data has provided valuable information about the mechanisms of protein–nucleic acid interactions. At the atomic level, the interactions involve a complex combination of hydrogen bonds, van der Waals contacts and water-mediated bonds between amino acid residues and nucleotide bases (4,5). It has also been found that the composition of DNA or RNA-binding residues is biased towards basic and polar amino acids (e.g. arginine and serine), while hydrophobic and acidic amino acids such as leucine and glutamic acid are statistically under-represented at the interaction interface (6–8). The information from structural analysis has been used to predict DNA-binding residues in solved protein structures (9,10).

However, it is still challenging to predict DNA or RNA-binding residues directly from amino acid sequence data, which are rapidly accumulating from many species. The problem for machine learning can be specified as follows: given the amino acid sequence of a protein that is supposed to bind DNA or RNA, the task is to predict which amino acid residues may be located at the interaction interface. Both the structure of the protein and the sequence of the target DNA or RNA are assumed to be unknown. Recently, artificial neural networks have been trained with sequence information and residue solvent accessibility for prediction of DNA-binding residues (8). The performance of the neural networks is at 40.3% sensitivity and 81.8% specificity. Evolutionary information in terms of a position-specific scoring matrix (PSSM) has been shown to enhance the predictive performance to 68.2% sensitivity and 66.0% specificity (11). To derive the PSSM, the query sequence is searched against a large reference database using the PSI-BLAST program (12). Thus, the PSSM-based method is computationally intensive and may not be used for efficient online predictions. Furthermore, PSSMs cannot be derived for query sequences that have no homologues in the reference database.

In the present work, support vector machines (SVMs) are trained using three simple sequence features for the prediction of DNA and RNA-binding residues. SVMs are a class of relatively new machine learning algorithms, which have recently been applied to a variety of biological problems for pattern classification (13). When compared with neural networks, SVMs may have advantages in their superior generalization power and the ability to avoid overfitting. We show that the SVM models can predict DNA-binding residues at 69.40% sensitivity and 70.47% specificity, while prediction of RNA-binding residues achieves 66.28% sensitivity and 69.84% specificity. Importantly, the three sequence features,

*To whom correspondence should be addressed. Tel: +1 785 532 6347; Fax: +1 785 532 6653; Email: ljwang@ksu.edu

including the side chain pK_a value, hydrophobicity index and molecular mass of an amino acid, are very efficient to compute for online predictions. Thus, we have developed the BindN web server (http://bioinformatics.ksu.edu/bindn/) for public access to the SVM classifiers.

MATERIALS AND METHODS

Datasets

Two amino acid sequence datasets, PDNA-62 and PRINR25, have been used to construct the SVM models for predicting DNA and RNA-binding residues, respectively. The PDNA-62 dataset was derived from 62 structures of representative protein–DNA complexes and had <25% identity among the sequences (8,11). The PRINR25 dataset was collected in this study from the protein–RNA complexes available at the Protein Data Bank (PDB, http://www.rcsb.org/pdb/). We selected 174 structures that had been determined by X-ray crystallography with resolution better than 3.5 Å. To remove redundancy among the amino acid sequences, clustering analysis was performed using the blastclust program in the BLAST package from NCBI (http://www.ncbi.nlm.nih.gov/BLAST/download.shtml). The blastclust program was run with the sequence identity threshold set to 25%, and the longest sequence in each cluster was selected for the non-redundant dataset, PRINR25.

As in the previous studies (8,11), an amino acid residue was designated as a binding site if the side chain or backbone atoms of the residue fell within a cutoff distance of 3.5 Å from any atoms of the DNA or RNA molecule in the complex. All the other residues were regarded as non-binding sites. A Perl program was developed to take a set of structure files as input and create an output file of amino acid sequences with each residue labeled as a binding or non-binding site according to the above criterion.

The lists of sequences in the PDNA-62 and PRINR25 datasets are provided in the supplementary material. The PDNA-62 dataset contains 1215 DNA-binding residues and 6948 non-binding residues, while the PRINR25 dataset has 3239 RNA-binding residues and 18 519 non-binding residues.

Training and testing

SVMs were trained with residue-wise data instances extracted from the sequence datasets. Each data instance was a subsequence of length w, where w was the sliding window size set to eleven in this study. Other window sizes were also tested, but the SVM classifiers constructed with $w = 11$ gave the best performance. From a protein sequence with n amino acid residues, a total of $(n - w + 1)$ data instances were extracted. The target residue was positioned in the middle of the subsequence, and the five neighboring residues on each side provided context information for the target residue. A data instance was labeled with 1 (positive) if the target residue was DNA or RNA-binding, or -1 (negative) if the target residue was non-binding.

Each residue was represented with three biochemical features, including the side chain pK_a value, hydrophobicity index and molecular mass of the amino acid. For a data instance with 11 residues, the input vector consists of 33 feature values. The biochemical features are very efficient to compute for a given amino acid sequence, and appear to be relevant for prediction of DNA and RNA-binding residues. The side chain pK_a value determines the ionization state of a residue. Since the phosphate groups of nucleic acids are negatively charged, the ionization state of amino acid side chains affects the interaction with DNA or RNA molecules. In this study, the side chain pK_a values from (14) were used, and the feature value was set to 7 for the amino acids with no side chain pK_a values. Hydrophobicity is a key factor in amino acid side chain packing and protein folding. Hydrophobic amino acids are often located inside globular proteins, but under-represented at the DNA or RNA interaction interfaces (6–8). We used the hydrophobicity scale developed by Kyte and Doolittle (15) to assign the feature values in this study. Since each amino acid has a unique value of molecular mass, this feature is used to represent the sequence information. Molecular mass is also related to the volume of space that a residue occupies in structures.

The SVMlight package (16) available at http://svmlight.joachims.org/ was used to construct the SVM classifiers. For a given set of binary-labeled training examples, SVM maps the input space into a higher-dimensional space and seeks a hyperplane to separate the positive data instances from the negative ones (17). The optimal hyperplane maximizes the separation margin between the two classes of training data, and is defined by a fraction of the input data instances close to the hyperplane (the so-called support vectors). The distance measurement between the data points in the high-dimensional space is defined by the kernel function. In this study, we used the radial basis function (RBF) kernel

$$K(\vec{x},\vec{y}) = \exp(-\gamma\|\vec{x}-\vec{y}\|^2), \quad \mathbf{1}$$

where $\vec{x}$ and $\vec{y}$ are two data vectors, and γ is a training parameter. A smaller γ value makes the decision boundary smoother. Another parameter for SVM training is the regularization factor C, which controls the trade-off between low training error and large margin (16). Different values for the C and γ parameters have been tested in this work to optimize the prediction accuracy.

A 5-fold cross-validation approach was used to evaluate the classifier performance. The positive and negative data instances were distributed randomly into five sets or the so-called folds. In each of the five iterative steps, four of the five sets were used to build a classifier (training), and then the classifier was evaluated using the remaining one set (testing). The predictions made for the test data instances in all the five iterations were combined and used to compute the results presented in this paper. However, the SVM models used by the BindN web server were constructed with all the available data instances.

Classifier performance measures

The predictions made for the test data instances are compared with the class labels (binding or non-binding) to evaluate the classifiers. The overall accuracy is defined as

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}, \quad \mathbf{2}$$

where TP is the number of true positives (binding residues with positive predictions); TN is the number of true negatives

(non-binding residues with negative predictions); FP is the number of false positives (non-binding residues but predicted as binding sites) and FN is the number of false negatives (binding residues but predicted as non-binding sites). However, the overall accuracy alone could be misleading in this case. Since there are more non-binding residues than binding ones in the datasets, a classifier can achieve >85% accuracy by simply predicting all the residues as negatives. Thus, sensitivity and specificity of the predictions are also computed as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \mathbf{3}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad \mathbf{4}$$

The receiver operating characteristic (ROC) curve is probably the most robust approach for classifier evaluation (18). The ROC curve is drawn by plotting the true positive rate (i.e. sensitivity) against the false positive rate, which equals to (1 − specificity). The area under the ROC curve (AUC) can be used as a reliable measure of classifier performance (19). Since the ROC plot is a unit square, the maximum value of AUC is 1, which is achieved by a perfect classifier. Weak classifiers and random guessing have AUC values close to 0.5.

RESULTS

Table 1 shows the performance of the SVM classifiers in 5-fold cross validations. The results have been obtained using the training parameters, $C = 0.5$ and $\gamma = 0.1$, which give better results than other values for prediction of both DNA and RNA-binding residues. The classifier for DNA-binding residues achieves 70.31% overall accuracy with 69.40% sensitivity and 70.47% specificity. For RNA-binding residues, predictions can be made at 69.32% overall accuracy with 66.28% sensitivity and 69.84% specificity (Table 1).

The ROC curves for prediction of DNA and RNA-binding residues are shown in Figure 1. These ROC curves have been generated by varying the output threshold of the SVM classifiers and plotting the true positive rate against false positive rate for each of the threshold values. The default output threshold used by SVMlight is 0, thus all the outputs ≥ 0 result in positive predictions and the outputs <0 give rise to negative predictions. When higher thresholds are used, only the data instances with relatively higher output values are predicted as positives and thus the true positive rate (sensitivity) becomes lower. Meanwhile, with higher thresholds, specificity becomes higher but the false positive rate (1 − specificity) gets lower. Therefore, each point on the ROC curve represents the trade-off between sensitivity and specificity. The ROC curves shown in Figure 1 are used by the BindN web server to allow users to specify the desired level of specificity or sensitivity (see below).

The ROC analysis shows that the classifier for DNA-binding residues is slightly more accurate than the classifier for RNA-binding residues, except at very low false positive rates (Figure 1). The AUC values are 0.7524 and 0.7308 for prediction of DNA and RNA-binding residues, respectively (Table 1). These AUC values are significantly higher than that of random guessing (0.5).

Table 1. Performance of the SVMs for prediction of DNA and RNA binding residues in proteins

Prediction type	Accuracy (%)	Sensitivity (%)	Specificity (%)	ROC AUC
DNA-binding	70.31	69.40	70.47	0.7524
RNA-binding	69.32	66.28	69.84	0.7308

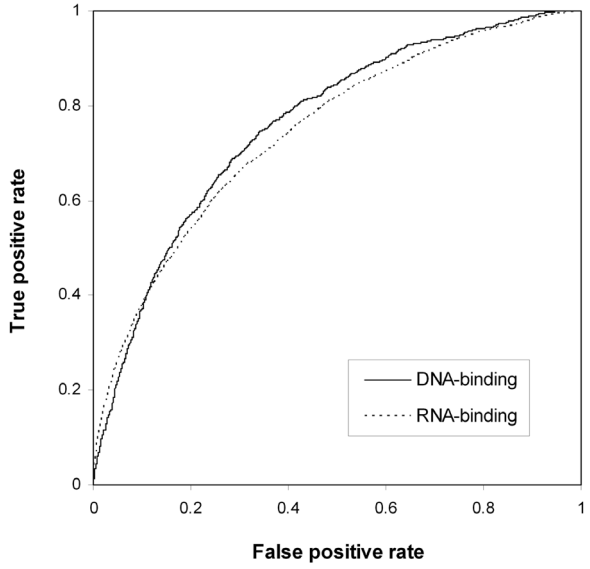


Figure 1. ROC curves for prediction of DNA and RNA-binding residues with SVMs.

The SVM classifier for DNA-binding residues appears to be better than the previous neural network predictors constructed using the same dataset (PDNA-62). The SVM's average of sensitivity and specificity is 69.94%, whereas the average (also called 'net prediction') was 61.1% for the neural network trained with sequence information and residue solvent accessibility (8). The PSSM-based approach improved the 'net prediction' to 67.1% (11), but required intensive computation for feature extraction. In contrast, the three sequence features used in our approach are very efficient to compute and thus well suited for online predictions.

For direct comparison of the SVM and neural network classifiers, a separate test dataset has been collected from the protein–DNA complexes available at PDB. As listed in Supplementary Table 3, the test dataset contains 92 amino acid sequences. These sequences show <30% identity among them and with any sequences in the PDNA-62 dataset. Except for the above constraints, the test dataset has been processed in the same way as described for the PRINR25 dataset. Putative DNA-binding residues have been predicted using both BindN and DBS-PSSM. The DBS-PSSM web server available at http://www.netasa.org/dbs-pssm/ was build using the PSSM-based neural network approach (11). The predictions have been made with the expected specificity at 72.3%, which is allowed at the DBS-PSSM server (DBS-PSSM does not allow users to specify their desired levels of specificity). As shown in Table 2, the actual specificity levels achieved by the two servers are close to the expected value. However, BindN achieves a much higher level of sensitivity than

DBS-PSSM (65.22 versus 36.73%). While the specificity achieved by BindN (65.22%) is close to the expected value based on the ROC analysis (67.19%), the actual specificity of DBS-PSSM does not reach the expected level (60.2%) on the new test dataset, probably owing to poor generalization of the representative DNA-binding residues in the relatively small training dataset (PDNA-62).

The SVM classifiers have been constructed using known DNA or RNA-binding proteins. The residues that do not bind to DNA or RNA have been used as the negative data instances for training. To further evaluate the SVM classifiers, we have analyzed a set of 100 proteins that do not interact with DNA or RNA. The protein sequences listed in Supplementary Table 4 have been selected randomly from the Swiss-Prot database (http://www.expasy.org/sprot/). When these sequences are analyzed using BindN with the expected specificity at 80% (default value), the actual specificity levels achieved by the SVM classifiers are 81.58 and 80.86% for the analysis of DNA and RNA-binding residues, respectively. The results suggest that BindN is reliable in achieving user-defined levels of specificity for various proteins. Thus, putative DNA or RNA-binding proteins with uncharacterized functional domains can be used as inputs to BindN. If the number of predicted binding residues is significantly higher than the expected number of false positives, the prediction results may be used to guide experimental characterization of these proteins.

Table 2. Performance comparison of the web servers for prediction of DNA-binding residues

Web server	Accuracy (%)	Sensitivity (%)	Specificity (%)
BindN	72.18	65.22	72.84
DBS-PSSM	67.82	36.73	70.79

To demonstrate that BindN can provide useful information for understanding protein–nucleic acid interactions, we have examined the predicted binding residues in the context of three-dimensional structures. Figure 2 shows two representative examples of the results. In Figure 2a, putative DNA-binding residues predicted by BindN for the mouse ETS-1 transcription factor are verified using the available structural data (PDB ID: 1K79). The structure includes residues 331–440 of the ETS-1 protein, which was not used for training the SVM classifier. The only homologue in the PDNA-62 dataset is the PU.1 DNA-binding domain (PDB ID: 1PUE), which has 28% sequence identity with the ETS-1 peptide. As shown in Figure 2a, 10 of the 16 DNA-binding residues (62.50%) are predicted correctly from the amino acid sequence data. These true positives are highlighted in red. The residues in blue are the six false negatives (DNA-binding residues but predicted as negatives). For the 88 non-binding residues, 79 or 89.77% are predicted correctly (residues in green), which agrees well with the desired level of specificity at 90%. Nevertheless, nine of the non-binding residues are predicted incorrectly (false positives in yellow). In Figure 2b, putative RNA-binding residues predicted for the archaeal protein L7Ae (box C/D RNA-binding domain) are examined. Chain B of the structure (PDB ID: 1RLG) is not included in the PRINR25 dataset,

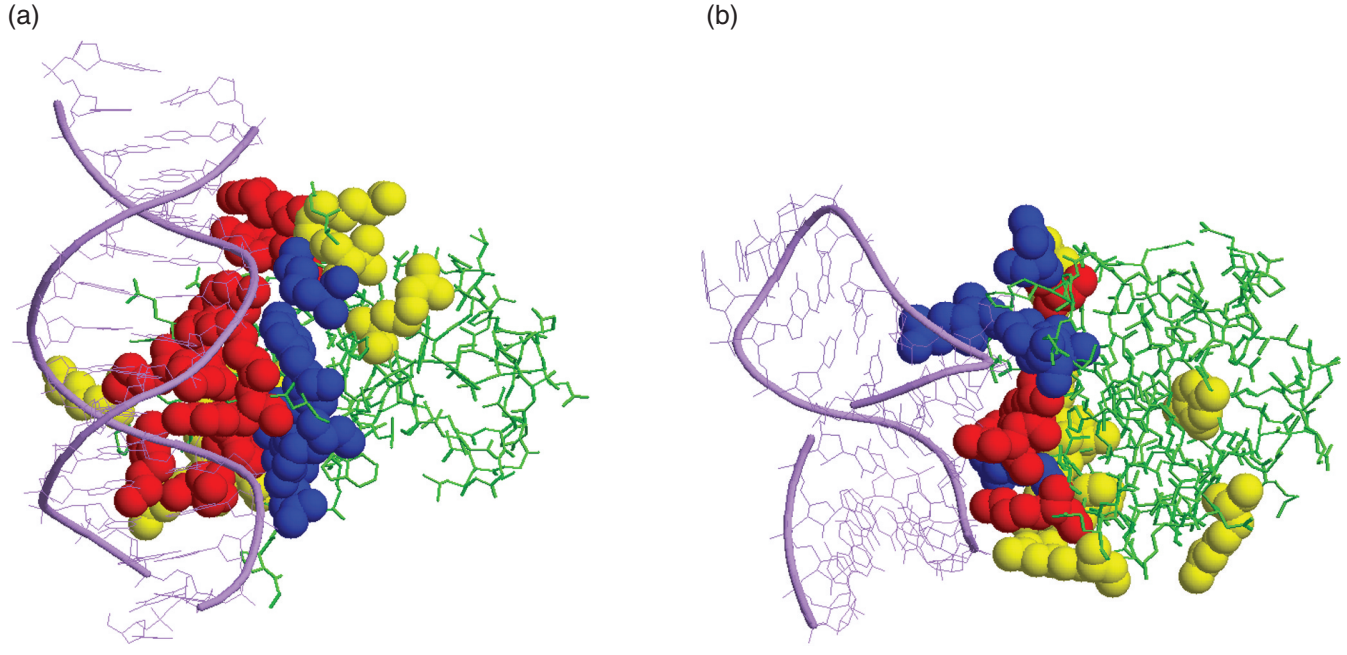


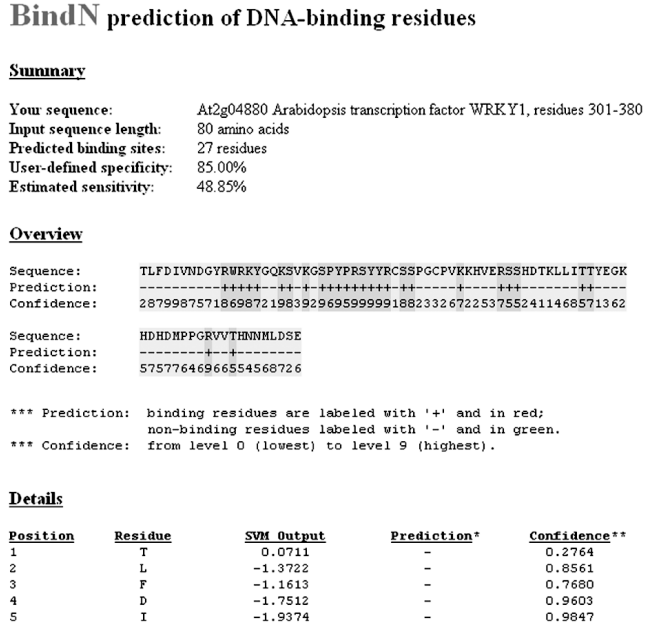
Figure 2. Representative prediction results shown in the context of three-dimensional structures. In each complex, the correctly predicted binding residues (true positives) are in red and spacefill; the correctly predicted non-binding residues (true negatives) are in green and wireframe; the binding residues but predicted as negatives (false negatives) are in blue and spacefill; the non-binding residues but predicted as positives (false positives) are in yellow and spacefill; the nucleic acid molecule is shown in purple. (**a**) Putative DNA-binding residues predicted for the mouse ETS-1 transcription factor. The structure (PDB ID: 1K79) includes residues 331–440 of the ETS-1 protein. Chain D of 1K79 was used as the input sequence to BindN with the expected specificity set to 90%. (**b**) Putative RNA-binding residues predicted for the box C/D RNA-binding domain of the archaeal protein L7Ae. Chain B of the structure (PDB ID: 1RLG) was used for BindN prediction with the expected specificity set to 90%.

but shows 34% sequence identity with the training data, 1E7K_A (human spliceosomal 15.5 KDa protein). Correct predictions are made for 6 of the 13 RNA-binding residues (46.15%) and 96 of the 104 non-binding residues (92.31%). Therefore, the predictions made by BindN can provide useful information for understanding protein–nucleic acid interactions. Such information may be used to guide experimental studies such as site-directed mutagenesis for functional characterization of DNA or RNA-binding proteins.

SERVER DESCRIPTION

On the BindN web page, users can enter an amino acid sequence in FASTA format; choose the type of prediction to be performed (either DNA or RNA-binding site prediction); and specify the desired level of sensitivity or specificity for the prediction results. The system uses the three biochemical features to encode the input sequence as described in Materials and Methods. The svm_classify program of the SVMlight package (16) is then called to classify the amino acid residues using one of the SVM models developed in this study. The user-defined level of sensitivity or specificity is used to choose the output threshold according to the ROC analysis discussed above. To reduce the number of false positive predictions, users may choose a specificity level higher than the default value (80%). In addition, if the DNA or RNA-binding domain is known, users may use the domain sequence instead of the full-length protein sequence as the input to BindN. Since a protein often has many more non-binding residues than binding ones, a long sequence, even with a high level of specificity, may result in a number of false positive predictions. However, the overall prediction accuracy appears to be at the same level for the binding domains and other sequence regions (data not shown).

As shown in Figure 3, the output report returned from the BindN server includes the following three sections: (i) a summary of the prediction results, including the number of binding sites predicted, the expected sensitivity and specificity; (ii) an overview of the predictions with the binding residues highlighted in red and the non-binding residues in green and (iii) detailed information about the prediction for each residue. For the summary, either sensitivity or specificity is chosen by the user, and the other value is estimated using the ROC curve shown in Figure 1. The overview can be used to examine the distribution of the predicted DNA or RNA-binding residues along the input sequence. In the output report, binding residues are labeled with '+' and non-binding residues are labeled with '−'. The confidence of prediction is based on the SVM output and computed as $(1 - s)$, where s is the expected sensitivity for positive predictions or specificity for negative predictions if the SVM output is used as the threshold in the ROC analysis. Intuitively, the confidence value indicates where the SVM output is ranked when compared with all the true positive or true negative predictions in the cross-validation. For example, the fifth residue (I) of the input sequence in Figure 3 gives the SVM output equal to −1.9374 and has the confidence for negative prediction equal to 0.9847, which indicates that only 1.53% of the non-binding residues in the PDNA-62 dataset have SVM outputs less than −1.9374. For the prediction overview, the confidence level is computed as the floor of (10 × confidence) so that it ranges from the lowest level 0 to the highest level 9 for presentation.



BindN prediction of DNA-binding residues

Summary

Your sequence:	At2g04880 Arabidopsis transcription factor WRKY1, residues 301-380
Input sequence length:	80 amino acids
Predicted binding sites:	27 residues
User-defined specificity:	85.00%
Estimated sensitivity:	48.85%

Overview

```
Sequence:    TLFDIVNDGYRWRKYGQKSVKGSPYPRSYYRCSSPGCPVKKHVERSSHDTKLLITTYEGK
Prediction:  ----------+++++--++-+-+++++++++-++-----+----+++-------++----
Confidence:  28799875718698721983929695999991882332672253755241146857136 2

Sequence:    HDHDMPPGRVVTHNNMLDSE
Prediction:  --------+--+--------
Confidence:  5757764696655456872 6
```

```
*** Prediction:  binding residues are labeled with '+' and in red;
                 non-binding residues labeled with '-' and in green.
*** Confidence:  from level 0 (lowest) to level 9 (highest).
```

Details

Position	Residue	SVM Output	Prediction*	Confidence**
1	T	0.0711	-	0.2764
2	L	-1.3722	-	0.8561
3	F	-1.1613	-	0.7680
4	D	-1.7512	-	0.9603
5	I	-1.9374	-	0.9847

Figure 3. Sample output from the BindN server. Putative DNA-binding residues were predicted for the *Arabidopsis* transcription factor WRKY1 (residues 301–380). Most of the positive predictions are located in the putative WRKY DNA-binding domain, for which structural data are still not available.

CONCLUSIONS

In this work, we have described a new SVM-based approach for prediction of DNA and RNA-binding residues based on amino acid sequence data. The average of sensitivity and specificity reaches nearly 70% for prediction of DNA-binding residues and ∼68% for prediction of RNA-binding residues. We have demonstrated that the prediction results can provide useful information for understanding protein–nucleic acid interactions. A web server, called BindN, has been developed for efficient online predictions. Currently, work is being performed to further improve the prediction accuracy by including more sequence features for the input encoding. In addition, the current system uses a fixed cutoff distance (3.5 Å) to discriminate binding residues from non-binding ones. This limitation will be overcome by constructing a series of SVM models with different cutoff distances and allowing users to choose which model to be used for online predictions.

SUPPLEMENTARY DATA

Supplementary data are available at *NAR* Online.

ACKNOWLEDGEMENTS

This work is supported by the K-INBRE Bioinformatics Core (NIH grant number P20 RR016475). Funding to pay the Open Access publication charges for this article was

provided by the K-INBRE Bioinformatics Core (NIH grant number P30 RR016475.

Conflict of interest statement. None declared.

REFERENCES

1. Ptashne,M. (2005) Regulation of transcription: from lambda to eukaryotes. *Trends Biochem. Sci*., **30**, 275–279.
2. Noller,H.F. (2005) RNA structure: reading the ribosome. *Science*, **309**, 1508–1514.
3. Hertel,K.J. and Graveley,B.R. (2005) RS domains contact the pre-mRNA throughout spliceosome assembly. *Trends Biochem. Sci*., **30**, 115–118.
4. Jones,S., Daley,D.T.A., Luscombe,N.M., Berman,H.M. and Thornton,J.M. (2001) Protein–RNA interactions: a structural analysis. *Nucleic Acids Res*., **29**, 943–954.
5. Luscombe,N.M., Laskowski,R.A. and Thornton,J.M. (2001) Amino acid-base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucleic Acids Res*., **29**, 2860–2874.
6. Nadassy,K., Wodak,S.J. and Janin,J. (1999) Structural features of protein–nucleic acid recognition sites. *Biochemistry*, **38**, 1999–2017.
7. Draper,D.E. (1999) Themes in RNA-protein recognition. *J. Mol. Biol*., **293**, 255–270.
8. Ahmad,S., Gromiha,M.M. and Sarai,A. (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.
9. Jones,S., Shanahan,H.P., Berman,H.M. and Thornton,J.M. (2003) Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res*., **31**, 7189–7198.
10. Tsuchiya,Y., Kinoshita,K. and Nakamura,H. (2005) PreDs: a server for predicting dsDNA-binding site on protein molecular surfaces. *Bioinformatics*, **21**, 1721–1723.
11. Ahmad,S. and Sarai,A. (2005) PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics*, **6**, 33.
12. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
13. Noble,W.S. (2004) Support vector machine applications in computational biology. In Scholkopf,B., Tsuda,K. and Vert,J.P. (eds), *Kernel Methods in Computational Biology*. MIT Press, Cambridge.
14. Nelson,D.L. and Cox,M.M. (2000) *Lehninger Principles of Biochemistry*. 3rd edn. Worth Publishers, New York.
15. Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol*., **157**, 105–132.
16. Joachims,T. (1999) Making large scale SVM learning practical. In Scholkopf,B., Burges,C. and Sola,A. (eds), *Advances in Kernel Methods—Support Vector Learning*. MIT Press, Cambridge.
17. Vapnik,V.N. (1998) *Statistical Learning Theory*. John Wiley and Sons, New York.
18. Swets,J.A. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.
19. Bradley,A.P. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, **30**, 1145–1159.